Total Path Variation for Deep Nets with General Activation Functions

August 11, 2019

Abstract

This paper shows that complexity bounds involving the total path variation (i.e., the "path norm") arise with any Lipschitz activation function which is zero at the origin. The heart of the analysis uses the probabilistic method to establish the existence of a sparse representer set for deep neural networks, which in turn, can be used to bound the metric entropy for a subcollection of all deep neural networks.

Index terms — Deep learning; neural networks; supervised learning; nonparametric regression; nonlinear regression; penalization; machine learning; high-dimensional data analysis; big data; statistical learning theory; generalization error; probabilistic method; variation; Markov chain; path norm; matrix product; quantization

1 Introduction

Statisticians and applied researchers are frequently concerned with predicting a response variable at a new input from a set of data collected from an experiment or observational study. We assume the learning (training) data is $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i)$, $1 \leq i \leq n$ are i.i.d. with common joint distribution $\mathbb{P}_{\mathbf{X},Y}$. Here, $\mathbf{X}_i \in \mathbb{R}^d$ is the feature (covariate vector) and $Y_i \in \mathbb{R}$ is a continuous outcome. A generic pair of variables will be denoted as $(\mathbf{X}, Y)$ with joint distribution $\mathbb{P}_{\mathbf{X},Y}$. A generic coordinate of $\mathbf{X}$ will be denoted by X. For convenience, we will often simply refer to X as a variable. We assume that $Y_i = f^*(\mathbf{X}_i) + \varepsilon_i$, for $i = 1, \dots, n$ (the statistical model) where $f^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ is an unknown regression function and $\{\varepsilon_i\}_{1 \leq i \leq n}$ are i.i.d. errors. The conditional average of Y given $\mathbf{X}$ is optimal in mean squared-error for the prediction of future Y from corresponding input X, if one uses squared error loss $L(Y, \widetilde{f}) = |Y - \widetilde{f}|^2$ since it minimizes the conditional risk $\mathbb{E}\left[|Y - \widetilde{f}(\mathbf{X})|^2 \mid \mathbf{X}\right] = \int |y - \widetilde{f}(\mathbf{X})|^2 \mathbb{P}_{Y|\mathbf{X}}(dy)$.

From the data, estimators $\hat{f}(\mathbf{x}) = \hat{f}(\mathbf{x}; \mathcal{D}_n)$ are formed. For concreteness, the loss at a target f^* is the $\mathbb{L}^2(\mathbb{P}_{\mathbf{X}})$ square error $\|f^* - \hat{f}\|^2 = \int |f^*(\mathbf{x}) - \hat{f}(\mathbf{x})|^2 \mathbb{P}_{\mathbf{X}}(d\mathbf{x})$ and the risk is the expected squared-error $\mathbb{E}\|f^* - \hat{f}\|^2$. For any class of functions $\mathcal{F}$ on $\mathbb{R}^d$, the minimax risk is

$$R_n(\mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\|f - \hat{f}\|^2\right], \tag{1}$$

where the infimum runs over all estimators $\hat{f}$ of f based on the data $\mathcal{D}_n$. We will investigate the behavior of $R_n(\mathcal{F})$ for deep neural network classes $\mathcal{F}$ [that are used to model a high-dimensional nonparametric regression function] and provide adaptive risk bounds for $\mathbb{E}\left[\|f^* - \hat{f}\|^2\right]$ when $\hat{f}$ is a obtained from complexity penalized empirical risk minimization. More specifically, a major focus of this article will be to investigate how $\mathbb{E}\left[\|f^* - \hat{f}\|^2\right]$ can be small even though n may be considerably smaller than the ambient dimension d or other parameters which define f^* [for example, if it is a deep neural network used to model a high-dimensional regression function].

2 Deep Learning Networks

2.1 Background and Prior Work

Good empirical performance of deep learning networks has been reported across various disciplines for difficult tasks in classification and prediction [LeCun et al., 2015]. These successes have largely been buoyed by the ability of multi-layer networks to generalize well despite being able to fit rich and complicated datasets, given enough parameters—an apparent contradiction to age-old statistical wisdom that warns against overfitting. This phenomenon is particularly striking when the input dimension is far greater than the available sample size, as is the case with many modern applications in molecular biology, medical imaging, and astrophysics, to name a few. Despite a vast amount of effort that goes into training deep learning models, typically in an ad-hoc manner for anecdotal datasets, a unifying theory of their complex mechanisms has not yet caught up with these applied and practical developments.

As is generally true in statistical estimation, there is a trade-off between estimation error and descriptive model complexity relative to sample size. Indeed, suppose $\hat{f}$ is a complexity penalized least squared estimator from data $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, over a class of candidate functions $\mathcal{F}$ [e.g., deep neural networks], i.e., $\hat{f}$ is chosen to optimize or approximately optimize

$$\frac{1}{n}\sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2 + \frac{\text{penalty}(f)}{n}, \tag{2}$$

over a collection $\mathcal{F}$ of candidate functions, where penalty(f) is a term that modulates some notion of the complexity of f. Then it is a classical fact [Barron et al., 1999, Barron, 1991] that $\hat{f}$ has the following adaptive risk bound:

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq C \inf_{f \in \mathcal{F}} \left\{ \|f^* - f\|^2 + \frac{\text{complexity}(f)}{n} \right\}, \tag{3}$$

where $C > 1$ is a universal constant and complexity(f) is a measure of descriptive complexity of a candidate fit f [typically complexity(f) is proportional to the δ-metric entropy of $\mathcal{F}$ at $\delta = \|f - f^*\|$]. The right side of (3) is an index of resolvability expressing the tradeoff between approximation error and descriptive complexity relative to sample size n. Thus, in analyzing the statistical properties of deep neural networks, in particular, one needs bounds on these two quantities. The forthcoming results attempt to address these aspects.

At the outset, one may be tempted to believe that the descriptive complexity of deep learning models is very large, in accordance with the large number of parameters that index each model. Fortunately, it is argued that, although a generic deep network may be difficult to describe, nevertheless, under suitable control on norms of the weights, it can be approximated well by a sparse representation, and this sparse representation comes from a subfamily that has a manageable cardinality. These small cardinality coverings can then be used to balance the estimation error and complexity trade-off [as per (3)] and thereby achieve [close to] optimal rates of estimation, in a minimax sense, in appropriate settings.

Prior results that seek to quantity different notions of model complexity typically produce unsavory statistical risk bounds for two main reasons.

First, the functions classes that are approximated by deep networks are typically not suited for high-dimensional settings. Indeed, minimax optimal rates for certain smooth function classes [e.g., Lipschitz, Hölder, Sobolev] degrade either with the number of inputs per layer, viz., $O(n^{-\alpha_d})$, where $\alpha_d \to 0$ as d approaches infinity, or in a similar way through the depth. Second, the complexity constants often scale exponentially with the depth or number of units per layer [Neyshabur et al., 2017, 2015, Golowich et al., 2017, Bartlett et al., 2017, Arora et al., 2018], which is problematic for high-dimensional or very deep networks. Indeed, many applications involve depths ranging from 2 or 3 to 22 [Szegedy et al., 2015] or, at the extreme end, 152 [He et al., 2016]. Furthermore, the input dimension d can be extremely large, possibly in the millions.

Other works [Yarotsky, 2017, 2018, Harvey et al., 2017] study the general approximation capabilities of deep networks using state-of-the-art VC dimension bounds $cTL\log(T/L) \leq \text{VCdim}(T, L) \leq CTL\log T$ for depth L ramp networks, where T is the number of weights. So the VC dimension is indeed linear in the number of parameters to within a log-factor. These results, again, when applied to a statistical learning setting, do not satisfactorily showcase the advantages of these model classes.

A main theme in this section aims to answer the following fundamental question: Assume that the target function f^* is equal to [or approximated well] by a deep network. For such a family, what is the size of the smallest subfamily with members that can approximate an arbitrary network within a desired level of accuracy?

Let $f(\mathbf{x}; \mathbf{W})$ be the parameterized family of depth L networks which map input vectors $\mathbf{x}$ of dimension d into output vectors of dimension d_{out}, where $f(\mathbf{x}; \mathbf{W})$ either takes the form $\mathbf{W}_1\phi(\mathbf{W}_2\phi(\cdots \mathbf{W}_{L-1}\phi(\mathbf{W}_L x)))$ or the form $\phi_{out}(\mathbf{W}_1\phi(\mathbf{W}_2\phi(\cdots \mathbf{W}_{L-1}\phi(\mathbf{W}_L x))))$, where ϕ_{out} is any Lipschitz(1) function, such as the fully-rectified linear function $\phi_{out}(z) = \text{sgn}(z)\min\{|z|, 1\}$, which is applied at the output, and ϕ is another Lipschitz(1) function, such as the positive-part activation function $\phi(z) = \max\{z, 0\}$ for scalar inputs z [also known as the ramp function or lower-rectified linear unit (ReLU)] applied at the internal layers. Note that our numbering scheme is the opposite of convention, where deeper layers are associated with smaller numbers and shallower layers are associated with higher numbers. However, it will be seen that our analysis is facilitated by such a labeling.

There are d_ℓ units on layer ℓ for $\ell = 0, 1, 2, \ldots, L$, with $d_0 = d_{out}$ on the outermost layer, and $d_L = d$ input units on the innermost layer, where, for analysis convenience, ℓ specifies the number of layers away from the output. It is typical practice to set $d_1, d_2, \ldots, d_{L-1}$ to be a common [possibly quite large] value h [known as the width], at least as large as arising from d. The units on layer ℓ are indexed by j_ℓ in $\{1, 2, \ldots, d_\ell\}$. Each $\mathbf{W}_\ell$ is the $d_{\ell-1} \times d_\ell$ matrix of weights and each matrix entry $w_{j_{\ell-1}, j_\ell} = \mathbf{W}_\ell[j_{\ell-1}, j_\ell]$ is the weight between unit $j_{\ell-1}$ in layer ℓ and unit j_ℓ in layer ℓ, where the index specifying layer ℓ is dropped when it is clear from the indices j_ℓ. Each coordinate of the input vector x is assumed to have a bounded range in $[-1, 1]$.

The focus of the present paper is on the case that ϕ_{out} is the identity and $d_{out} = 1$, though multidimensional outputs can be examined similarly. In this case, there is but one output index $j_0 = 1$, and $\mathbf{W}_1$ is a row vector of length d_1 with entries $w_{j_0,j_1} = w_{j_1}$. Accordingly, for networks of the first form, the function $f(\mathbf{x};\mathbf{W})$ is

$$\sum_{j_1} w_{j_1}\phi(\sum_{j_2} w_{j_1,j_2}\phi(\sum_{j_3} w_{j_2,j_3}\cdots\phi(\sum_{j_L} w_{j_{L-1},j_L}x_{j_L}))). \tag{4}$$

Each unit computes $x_{j_\ell} = \phi(\sum_{j_{\ell+1}} w_{j_\ell,j_{\ell+1}}x_{j_{\ell+1}})$, where x_{j_ℓ} denotes the output value for unit j_ℓ on layer ℓ, as a function of its inputs $x_{j_{\ell+1}}$, starting with the innermost layer.

In [Barron and Klusowski, 2018], the authors examine the statistical risk [mean squared predictive error] of multi-layer networks with ℓ^1-type controls on their parameters and with ramp activation functions [also called lower-rectified linear units ReLU]. In this setting, the mean-squared predictive error [$\mathbb{L}^2$ risk] of an ℓ^1-type complexity regularized estimator was shown to be upper bounded by $[(L^3\log d)/n]^{1/2}$, where d is the input dimension to each layer, L is the number of layers, and n is the sample size. Similar bounds hold for generalization error and Rademacher complexity. In this way, the input dimension can be much larger than the sample size and the estimator can still be accurate, provided the target function has such ℓ^1 controls and that the sample size is at least moderately large compared to $L^3\log d$. The heart of the analysis is in the development of a sampling strategy that demonstrates the accuracy of a sparse covering of deep ramp networks.

Theorem 1 ([Barron and Klusowski, 2018]). *Consider the parameterized family $\mathcal{F}(L,\mathscr{V})$ of depth L ReLU networks with composite variation V at most $\mathscr{V}$.*[1] *There is a subfamily $\widetilde{\mathcal{F}}_M$ with log-cardinality of order*

$$(L-2)M\log(\max_{2\le\ell\le L}\min\{M,d_\ell\}) + M\log d,$$

such that for any probability measure $\mathbb{P}$ on $[-1,1]^d$ and any $f(\mathbf{x};\mathbf{W})$ belonging to $\mathcal{F}(L,\mathscr{V})$, there is a sparse approximant $f(\mathbf{x};\widetilde{\mathbf{W}})$ in $\widetilde{\mathcal{F}}_M$, with at most LM nonzero weights, such that

$$\int |f(\mathbf{x};\widetilde{\mathbf{W}}) - f(x;\mathbf{W})|^2\mathbb{P}(d\mathbf{x}) \le \left[\frac{L\mathscr{V}}{\sqrt{M}}\right]^2.$$

Recent related work has focused on using accuracy statements such as Theorem 1 to bound the Rademacher complexity of various deep network classes and studying how they can be used to bound the generalization error [i.e., the difference between the empirical error on the training dataset and the expected error on the underlying joint probability distribution with respect to some loss function] [Neyshabur et al., 2017, 2015, Golowich et al., 2017, Bartlett et al., 2017, Arora et al., 2018]. Typical results are of the following form: given access to a sample of size n drawn from the network, the generalization error scales as $\mathcal{C}/\sqrt{n}$, where $\mathcal{C}$ is some complexity constant that depends on the parameters of the network. In all these works, however, $\mathcal{C}$ has exponential dependence on L, either *indirectly* through some product of norms $\prod_{\ell=1}^{L}\|\mathbf{W}_\ell\|$ of the individual weight matrices $\mathbf{W}_\ell$ across the layers $\ell = 1,2,\dots,L$,

[1] A complexity constant that involves only the entry-wise ℓ^1 norms of successive products of the weight matrices, i.e., $\||\mathbf{W}_1||\mathbf{W}_2|\cdots|\mathbf{W}_\ell|\|_1$ and $\||\mathbf{W}_{\ell+1}||\mathbf{W}_2|\cdots|\mathbf{W}_L|\|_1$. [For a matrix A, $\|A\|_1 = \sum_{j_1,j_2}|A[j_1,j_2]|$]. See [Barron and Klusowski, 2018] for further details.

or *directly* as c^L for some positive constant $c > 1$ [in addition to polynomial factors in L or logarithmic factors in $d_1, d_2, \ldots, d_L$]. Notably, the work of [Golowich et al., 2017] manages to avoid this direct dependence on L by controlling various Schatten norms [The r-Schatten norm of a matrix is the ℓ^r norm of its singular values.] of the weight matrices.

The form $\prod_{\ell=1}^{L} \|\mathbf{W}_\ell\|$ of the complexity constants is an artifact of Rademacher analysis, which necessitates applying some sort of sub-multiplicative matrix norm inequality at the "peeling" step [whereby the complexity bound is inductively reduced to a complexity bound involving shallower networks]. Our probabilistic method avoids these reductions and instead works with all the weights at once.

We will now discuss some reasons to prefer the complexity constant V in Theorem 1 over other complexity constants in the literature. First, V involves *norms of matrix products* of the weight matrices, differs from complexity constants involving the *product of individual matrix norms* of the weight matrices, i.e., $\prod_{\ell=1}^{L} \|\mathbf{W}_\ell\|$, [Neyshabur et al., 2017, Golowich et al., 2017, Bartlett et al., 2017]. To see the utility of working with V, consider the simple case when all the weight matrices are the same and equal to a $d \times d$ matrix Q. Then $\|\mathbf{W}_1 \cdots \mathbf{W}_L\|_1$ is bounded by a constant factor, independent of L, times $[\rho(Q)]^L$, where $\rho(Q)$ is the spectral radius of Q. On the other hand, $\prod_{\ell=1}^{L} \|\mathbf{W}_\ell\| = \|Q\|^L \geq [\rho(Q)]^L$. Hence if $\rho(Q) \leq 1$, while $\|\mathbf{W}_\ell\| > 1$, the growth of the two complexity constants could be, at the extreme, the difference between constant and exponential in L. This could be problematic since some applications involve very large depths, e.g., $L = 152$ in [He et al., 2016]. Second, because any two matrix norms are equivalent, the ℓ^1 norm of a product of matrices can be bounded by the product of other individual matrix norms, provided they are submultiplicative. Taken together, these facts imply that the product of the weight matrices is a fundamental quantitative measure of complexity, from which other complexity constants in the literature can be obtained.

Yet another reason why the ℓ^1 norm of a matrix product is a natural complexity constant [if the input is contained in an ℓ^∞ ball] is that the risk bound (7) is akin to those obtained in [Raskutti et al., 2011, Theorem 4] or [Rigollet and Tsybakov, 2011, Theorem 3.2] for squared-error prediction in high-dimensional linear regression with ℓ^1 controls on the parameter vectors. To highlight the relationship between the linear and nonlinear case, if one instead used a linear activation function $\phi(z) = z$, the functions (4) would be deep linear networks [Ji and Telgarsky, 2018]

$$f(\mathbf{x}; \mathbf{W}) = \mathbf{W}_1 \cdots \mathbf{W}_L \mathbf{x}. \tag{5}$$

In this case, an ℓ^1 control on $\mathbf{W}_1 \cdots \mathbf{W}_L$, say $V = \|\mathbf{W}_1 \cdots \mathbf{W}_L\|_1 \leq \mathscr{V}$, also leads to squared-error prediction of order $\mathscr{V}\left(\frac{\log d}{n}\right)^{1/2}$. Of course, there is an important difference here—the richness of $\mathcal{F}(L, \mathscr{V})$ is determined by the expressiveness afforded by the nonlinearities and also by the variation through v. Therefore $\mathcal{F}(L, \mathscr{V})$ more flexibly represents a larger class of functions, far beyond the rigidity of linear.

Statements in the same style as Theorem 1 can be translated into bounds on the minimax risk (1), as will now be discussed. If $\mathcal{F}(L, \mathscr{V})$ denotes the collection of all such depth L networks with $V(f) \leq \mathscr{V}$ and $\mathbb{L}^\infty$ norm bounded by a constant, then Theorem 1 shows that the $\mathbb{L}^2$ ϵ-covering entropy of $\mathcal{F}(L, \mathscr{V})$, denoted by $\mathcal{V}_{\mathcal{F}(L,\mathscr{V})}(\epsilon)$, is of order

$$\frac{L^3 \mathscr{V}^2 \log(\max_\ell d_\ell)}{\epsilon^2}. \tag{6}$$

[Recall the definition of ϵ-covering entropy: Let $\mathbb{P}$ be a probability measure on a measurable space and suppose $\mathcal{F}$ is a family of functions in $\mathbb{L}^2(\mathbb{P})$. A subfamily $\widetilde{\mathcal{F}}$ is called an ϵ-covering for $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists $\widetilde{f} \in \widetilde{\mathcal{F}}$ such that $\|f - \widetilde{f}\| \leq \epsilon$. The logarithm of the minimum cardinality of ϵ-nets is called the ϵ-covering entropy of $\mathcal{F}$ and is denoted by $\mathcal{V}_{\mathcal{F}}(\epsilon)$.] For Gaussian errors $\varepsilon = Y - f(X)$, [Yang and Barron, 1999] show that the minimax risk is essentially governed by ϵ_n^2, where $\mathcal{V}_{\mathcal{F}}(\epsilon_n) \asymp n\epsilon_n^2$. Thus, one can deduce the following result from [Yang and Barron, 1999] and the ϵ-covering entropy bound (6) by solving $\mathcal{V}_{\mathcal{F}}(\epsilon_n) \asymp n\epsilon_n^2$ for the function class $\mathcal{F} = \mathcal{F}(L, \mathscr{V})$. Thus, the minimax risk bound (1) becomes

$$R_n(\mathcal{F}(L, \mathscr{V})) \leq C\mathscr{V} \left(\frac{L^3 \log(\max_{\ell \geq 2} d_\ell)}{n}\right)^{1/2}, \tag{7}$$

for some universal positive constant $C > 0$. This risk bound is surprising since it shows that the effect of large depth L and interlayer dimensions $d_1, d_2, \dots, d_L$ is relatively harmless and benign, even for modest sample sizes. This may explain why the performance of deep networks does not seem to be hindered by their highly parameterized structure. Also important to notice is that the rate in the exponent (1/2) does not degrade with the input dimension or other network parameters. Indeed, the main dependence on the target regression function f^* [which in this case is a deep neural network] is through a complexity constant v that depends only on products of its weight matrices $\mathbf{W}_1^*, \dots, \mathbf{W}_L^*$ and a low order polynomial in the depth L.

3 Main Results

We now turn our attention to the main topic of this paper. It is natural to ask whether there a general theory for quantifying the size of a sparse covering of deep neural networks for any activation function, not just ReLU.

Choosing the right activation function depends on the problem at hand; there is no single foolproof activation or loss function which yields ideal results in all the models. For example, the ReLU function suffers from the so-called "dying ReLU problem" [so that large numbers of units are pushed into inactive states for all inputs, thereby lowering the model capacity]. This prompts users to try other activation functions such as, for example, softplus $\phi(z) = \log(1 + e^z)$ [Glorot et al., 2011] or hyperbolic tangent $\phi(z) = [e^z - e^{-z}]/[e^z + e^{-z}]$.

As mentioned previously, it is desirable to have the complexity constants in the accuracy bounds involve norms of products of the weight matrices. Theorem 1 shows this type of result for ℓ^1 norms of products of the weight matrices $\mathbf{W}_1, \dots, \mathbf{W}_L$ for ReLU activation functions $\phi(z) = \max\{0, z\}$ [Barron and Klusowski, 2018]. However, the analysis hinges crucially on the positive homogeneity of the ReLU [i.e., $\phi(rz) = r\phi(z)$ for $r \geq 0$], and so a completely new technique must be developed if there is any hope in generalizing the results for non-homogenous activations that are popular alternatives when the ReLU is not justifiable for the particular data setting. Towards the direction, assume henceforth that ϕ is a general Lipschitz(1) activation function with $\phi(0) = 0$.

3.1 Probabilistic Method

We will use a trick [known as the probabilistic method] in which, for any collection of L weight matrices $\mathbf{W}$, representer parameters $\widetilde{\mathbf{W}}$ are drawn at random from a finite pre-specified set and then it is shown that the desired accuracy bound holds for the expectation, so accordingly there exists a representer of that accuracy. This type of reasoning in the context of function approximation is due to Pisier and Maurey [Pisier, 1980-1981] and was later applied to nonparametric regression with single-hidden layer networks in the seminal work of Barron [Barron, 1993, 1991].

The probabilistic method in the current setting has the following schema.

1. Let $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_1, \dots, \widetilde{\mathbf{W}}_L)$ be representer weights, drawn at random from a finite prespecified set $\widetilde{\mathbf{W}}$. The representer set is indexed by a parameter M that controls both the accuracy and the cardinality.

2. Suppose $\mathrm{E}_{\widetilde{\mathbf{W}}}[\|f(\mathbf{x};\widetilde{\mathbf{W}}) - f(\mathbf{x};\mathbf{W})\|] \leq \delta_M$. Then there exists a realization $\widetilde{\mathbf{W}}'$ of $\widetilde{\mathbf{W}}$ such that $\|f(\mathbf{x};\widetilde{\mathbf{W}}') - f(\mathbf{x};\mathbf{W})\| \leq \delta_M$.

3. The set of all possible realizations of $\widetilde{\mathbf{W}}$, mainly $\{f(\mathbf{x};\widetilde{\mathbf{W}}) : \widetilde{\mathbf{W}} \in \widetilde{\mathbf{W}}\}$, forms a δ_M-cover for $\{f(\mathbf{x};\mathbf{W}) : \mathbf{W} \in \mathbf{W}\}$, where $\mathbf{W}$ is a collection of sequences of L weight matrices.

4. If a typical realization of $\widetilde{\mathbf{W}}$ is sparse [as controlled by M], then the cardinality of $\widetilde{\mathbf{W}}$ will be small.

The bulk of the forthcoming work will be in establishing the second step, i.e., that $\mathrm{E}_{\widetilde{\mathbf{W}}}[\|f(\mathbf{x};\widetilde{\mathbf{W}}) - f(\mathbf{x};\mathbf{W})\|] \leq \delta_M$. This necessitates specifying a distribution on the weights, which we will now do.

Working with $\pm\phi$ and doubling the number of weights per layer, it can be assumed that the weights $w_{j_{\ell-1},j_\ell}$ are nonnegative, and indeed, we do so for the rest of the discussion. We first define a joint distribution across the indices $(j_1, \dots, j_L)$. This joint distribution has conditionals, for index j_ℓ given $j_{\ell-1}$, defined as

$$p_{j_\ell|j_{\ell-1}} = \frac{w_{j_{\ell-1},j_\ell}}{v_{j_{\ell-1}}},$$

where $v_{j_{\ell-1}} = \sum_{j_\ell} w_{j_{\ell-1},j_\ell}$ and $v_{j_0} = \sum_{j_1} w_{j_1}$. For the moment, assume the v_{j_ℓ} are known; later on, they will be replaced with approximants $\widetilde{v}_{j_\ell}$.

This setup facilitates a probabilistic interpretation of a deep neural network (4) as an iterated expectation interspersed with nonlinearities:

$$f(\mathbf{x};\mathbf{W}) = \sum_{j_1} v_{j_0} p_{j_1} \phi\big(v_{j_1} \sum_{j_2} p_{j_2|j_1} \phi\big(v_{j_2} \sum_{j_3} p_{j_3|j_2} \cdots \phi\big(v_{j_{L-1}} \sum_{j_L} p_{j_L|j_{L-1}} x_{j_L}\big)\big)\big).$$

Note that this representation is different from [Barron and Klusowski, 2018], where it was shown that for ReLU networks,

$$f(\mathbf{x};\mathbf{W}) = V \sum_{j_1} p_{j_1} \phi\big(\sum_{j_2} p_{j_2|j_1} \phi\big(\sum_{j_3} p_{j_3|j_2} \cdots \phi\big(\sum_{j_L} p_{j_L|j_{L-1}} x_{j_L}\big)\big)\big),$$

and $V = \|\mathbf{W}_1 \mathbf{W}_2 \cdots \mathbf{W}_L\|_1$.

In forming the [random] approximant $f(\mathbf{x};\widetilde{\mathbf{W}})$, take a random sample of size M from the joint distribution defined by

$$p_{j_1,j_2,\dots,j_L} = p_{j_1} p_{j_2|j_1} \cdots p_{j_L|j_{L-1}}. \tag{8}$$

Let $K_{j_1,j_2,\ldots,j_L} \sim \text{Multinomial}(M, (p_{j_1,j_2,\ldots,j_L}))$ be the corresponding empirical counts of occurrences of $(j_1, j_2, \ldots, j_L)$, distributed as multinomial, with empirical marginal counts $K_{j_{\ell_1},j_{\ell_2},\ldots,j_{\ell_k}}$ formed by summing over all unspecified indices. Representers of the original scaled network weights p_{j_ℓ} and $p_{j_\ell|j_{\ell-1}}$ are formed by taking

$$\widetilde{p}_{j_\ell} = \frac{K_{j_\ell}}{M} \quad \text{and} \quad \widetilde{p}_{j_\ell|j_{\ell-1}} = \frac{K_{j_{\ell-1},j_\ell}}{K_{j_{\ell-1}}},$$

with the convention that $0/0 = 0$.[2]

Each unit of the network computes $\widetilde{x}_{j_{\ell-1}} = \phi\big(\sum_{j_\ell} \widetilde{x}_{j_\ell}\widetilde{w}_{j_{\ell-1},j_\ell}\big)$, with $\widetilde{w}_{j_{\ell-1},j_\ell} = v_{j_{\ell-1}}\widetilde{p}_{j_\ell|j_{\ell-1}}$, and the full approximant takes on the form

$$f(\mathbf{x}; \widetilde{\mathbf{W}}) = \sum_{j_1} \widetilde{w}_{j_1}\phi\big(\sum_{j_2} \widetilde{w}_{j_1,j_2}\phi\big(\sum_{j_3} \widetilde{w}_{j_2,j_3} \cdots \phi\big(\sum_{j_L} \widetilde{w}_{j_{L-1},j_L} x_{j_L}\big)\big)\big). \tag{9}$$

Note that by construction, each $\widetilde{w}_{j_{\ell-1},j_\ell}$ is an unbiased estimate of $w_{j_{\ell-1},j_\ell}$.

3.2 Cardinality of Cover

Each $f(\mathbf{x}; \widetilde{\mathbf{W}})$ is built from empirical counts $K_{j_1,j_2,\ldots,j_L}$ of specified sum M. The number of ways one can have $\sum_{j_1,j_2,\ldots,j_L} K_{j_1,j_2,\ldots,j_L} = M$ is equal to

$$\binom{d_1 d_2 \cdots d_L + M - 1}{M}, \tag{10}$$

[by Feller's stars-and-bars argument [Feller, 1971, page 38]] with log-cardinality bounded by $M \log(2e\, d_1 d_2 \cdots d_L/M)$ whenever $d_1 d_2 \cdots d_L > M - 1$. Furthermore, as explained in [Barron and Klusowski, 2018, Section 5] when $d_\ell \geq M$, we may replace d_ℓ, $\ell = 1, 2, \ldots, L-1$, with $d_\ell^{new} = \min\{d_\ell, M\}$ in representation of $f(\mathbf{x}; \widetilde{\mathbf{W}})$. As such, the set of different realizations of $f(\mathbf{x}; \widetilde{\mathbf{W}})$ has manageable cardinality of order $LM \log(\max_{\ell \geq 2} d_\ell)$.

3.3 Bounding the Accuracy

To analyze $\mathrm{E}_{\widetilde{\mathbf{W}}}|f(\mathbf{x}; \widetilde{\mathbf{W}}) - f(\mathbf{x}; \mathbf{W})|^2$, we first write the difference between $f(\mathbf{x}; \widetilde{\mathbf{W}})$ and $f(\mathbf{x}; \mathbf{W})$ as a telescoping sum of [successively collapsing] differences $f(\mathbf{x}; \widetilde{\mathbf{W}}) - f(\mathbf{x}; \mathbf{W}) = \sum_{\ell=0}^{L-1}[f_{\ell+1}(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W}) - f_\ell(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W})]$ in which the $f_{\ell+1}(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W})$ and $f_\ell(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W})$ differ only on layer $\ell + 1$, the former using $\widetilde{w}_{j_\ell,j_{\ell+1}}$ and the later using $w_{j_\ell,j_{\ell+1}}$, i.e.,

$$f_{\ell+1}(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W}) = f(\mathbf{x}; (\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_\ell, \widetilde{\mathbf{W}}_{\ell+1}, \mathbf{W}_{\ell+2}, \ldots, \mathbf{W}_L)),$$

and

$$f_\ell(\mathbf{x}; \widetilde{\mathbf{W}}, \mathbf{W}) = f(\mathbf{x}; (\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_\ell, \mathbf{W}_{\ell+1}, \mathbf{W}_{\ell+2}, \ldots, \mathbf{W}_L)),$$

respectively.

[2] Note that if $(M_1, M_2, M_3) \sim \text{Multinomial}(M, (p_1, p_2, p_3))$, then $M_1/(M_1 + M_2)$ is the maximum likelihood estimator of $p_1/(p_1 + p_2)$. Analogously, since $(K_{j_{\ell-1},j_\ell}, K_{j_{\ell-1}} - K_{j_{\ell-1},j_\ell}, M - K_{j_{\ell-1}}) \sim \text{Multinomial}(M, (p_{j_{\ell-1},j_\ell}, p_{j_{\ell-1}} - p_{j_{\ell-1},j_\ell}, 1 - p_{j_{\ell-1}}))$, it follows that $K_{j_{\ell-1},j_\ell}/K_{j_{\ell-1}}$ is the maximum likelihood estimator of $p_{j_\ell|j_{\ell-1}}$.

By the triangle inequality, one can bound $\mathrm{E}[\int |f(\mathbf{x};\widetilde{\mathbf{W}})-f(\mathbf{x};\mathbf{W})|\mathbb{P}(d\mathbf{x})]$ by bounding each $\mathrm{E}[\int |f_{\ell+1}(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W}) - f_{\ell}(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W})|\mathbb{P}(d\mathbf{x})]$ and summing from $\ell = 0$ to $\ell = L-1$. Repeated application of the Lipschitz property of ϕ permits bounding each difference $|f_{\ell+1}(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W}) - f_{\ell}(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W})|$ by

$$\sum_{j_1,\dots j_{\ell-1},j_\ell} \widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|\phi(z_{j_\ell})-\phi(\widetilde{z}_{j_\ell})|, \tag{11}$$

where $z_{j_\ell} = \sum_{j_{\ell+1}} w_{j_\ell,j_{\ell+1}}x_{j_{\ell+1}}$ and $\widetilde{z}_{j_\ell} = \sum_{j_{\ell+1}} \widetilde{w}_{j_\ell,j_{\ell+1}}x_{j_{\ell+1}}$. Note that $\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}$ and $\widetilde{z}_{j_\ell}$ are conditionally independent given K_{j_ℓ} and hence,

$$\begin{aligned}&\mathrm{E}[\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|\phi(\widetilde{z}_{j_\ell})-\phi(z_{j_\ell})||K_{j_\ell}]\\ &= \mathrm{E}[\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|K_{j_\ell}]\mathrm{E}[|\phi(\widetilde{z}_{j_\ell})-\phi(z_{j_\ell})||K_{j_\ell}].\end{aligned}$$

The following identity, which is based on the fact that the $K_{j_{\ell-1},j_\ell}$ are conditionally binomially distributed given K_{j_ℓ}, reveals how the product of the path weights arise after taking iterated expectations [conditioning successively on $K_{j'_\ell}$] with respect to $\widetilde{\mathbf{W}}$:

$$\begin{aligned}\mathrm{E}[\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|K_{j_\ell}] &= v_{j_0}v_{j_1}p_{j_1|j_2}\cdots v_{j_{\ell-1}}p_{j_{\ell-1}|j_\ell}\frac{K_{j_\ell}}{M}\\ &= v_{j_0}p_{j_1}v_{j_1}p_{j_2|j_1}\cdots v_{j_{\ell-1}}p_{j_\ell|j_{\ell-1}}\frac{K_{j_\ell}}{p_{j_\ell}M}\\ &= w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}\frac{K_{j_\ell}}{p_{j_\ell}M},\end{aligned} \tag{12}$$

where $p_{j_{\ell-1}|j_\ell} = \frac{p_{j_{\ell-1},j_\ell}}{p_{j_\ell}} = \frac{p_{j_{\ell-1}}}{p_{j_\ell}}p_{j_\ell|j_{\ell-1}}$ are the reverse [backward] conditionals and p_{j_ℓ} is the marginal probability of the index j_ℓ, which can be expressed as

$$p_{j_\ell} = \sum_{j_1,j_2,\dots,j_{\ell-1}} \frac{w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}}{v_{j_0}v_{j_1}\cdots v_{j_{\ell-1}}} = \sum_{j_1,j_2,\dots,j_{\ell-1}} \frac{w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}}{\sum_{j'_1} w_{j'_1}\sum_{j'_2} w_{j_1,j'_2}\cdots\sum_{j'_\ell} w_{j_{\ell-1},j'_\ell}}$$

($\geq \min_{j_{\ell-1}} \frac{w_{j_{\ell-1},j_\ell}}{v_{j_{\ell-1}}}$) by summing out all indices other than j_ℓ in (8).

Using a similar analysis in [Barron and Klusowski, 2018], the conditional expected value of the difference $|\phi(\widetilde{z}_{j_\ell})-\phi(z_{j_\ell})|$ given K_{j_ℓ} is at most $v_{j_\ell}\frac{\sigma_{j_\ell}(\mathbf{x})}{\sqrt{K_{j_\ell}}}$, where $\mu_{j_\ell} = \sum_{j_{\ell+1}} p_{j_{\ell+1}|j_\ell}x_{j_{\ell+1}}$ and $\sigma^2_{j_\ell}(\mathbf{x}) = \sum_{j_{\ell+1}} p_{j_{\ell+1}|j_\ell}(x_{j_{\ell+1}} - \mu_{j_\ell})^2$ are the mean and variance, respectively, of $z_{\widetilde{j}}$ resulting from a single draw $\widetilde{j} \sim p_{j_{\ell+1}|j_\ell}$. Applying this bound to the expected value of (11) and using the identity from (12), one can bound the expected difference $|f_{\ell+1}(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W}) - f_\ell(\mathbf{x};\widetilde{\mathbf{W}},\mathbf{W})|$ by

$$\frac{1}{\sqrt{M}}\sum_{j_1,j_2,\dots,j_\ell,j_{\ell+1}} w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}w_{j_\ell,j_{\ell+1}}\sigma_{j_\ell}\sqrt{1/p_{j_\ell}},$$

or equivalently,

$$\frac{1}{\sqrt{M}}\|\mathbf{W}_1\mathbf{W}_2\cdots\mathbf{W}_\ell\mathbf{W}'_{\ell+1}\|_1, \tag{13}$$

if $\mathbf{W}'_{\ell+1}[j_\ell,j_{\ell+1}]$ is defined as $w_{j_\ell,j_{\ell+1}}\sigma_{j_\ell}\sqrt{1/p_{j_\ell}}$ if $p_{j_\ell} > 0$ and zero otherwise and $\sigma^2_{j_\ell} = \int \sigma^2_{j_\ell}(\mathbf{x})\mathbb{P}(d\mathbf{x})$. A final bound on entire expected difference $\mathrm{E}[\int |f(\mathbf{x};\widetilde{\mathbf{W}}) - f(\mathbf{x};\mathbf{W})|\mathbb{P}(d\mathbf{x})]$ results from summing (13) from $\ell = 0$ to $\ell = L-1$. This motivates the following definition.

Definition 1. *Let $f(\mathbf{x};\mathbf{W})$ be a depth L network with weight matrices $\mathbf{W} = (\mathbf{W}_1, \mathbf{W}_2, \dots, \mathbf{W}_L)$. The average path variation $V = V(f)$ is defined by*

$$\frac{1}{L}\sum_{\ell=0}^{L-1}\sum_{j_1,j_2,\dots,j_\ell,j_{\ell+1}} w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1}j_\ell}w'_{j_\ell,j_{\ell+1}} = \frac{1}{L}\sum_{\ell=0}^{L-1}\|\mathbf{W}_1\mathbf{W}_2\cdots\mathbf{W}_\ell\mathbf{W}'_{\ell+1}\|_1. \tag{14}$$

Remark 1. *The complexity bound also has the following upper bound, which is similar to (and improves upon) [Golowich et al., 2017, Theorem 2], which is purely in terms of products of $\ell_{1,\infty}$ norms of the individual weight matrices, $\|\mathbf{W}_1\|_1\|\mathbf{W}_2\|_{1,\infty}\cdots\|\mathbf{W}_L\|_{1,\infty}$,*

$$V \le \frac{1}{L}\sum_{\ell=0}^{L-1}\sqrt{\|\mathbf{W}_1\|_1\|\mathbf{W}_2\|_{1,\infty}\cdots\|\mathbf{W}_\ell\|_{1,\infty}}\sum_{j_\ell,j_{\ell+1}}\sigma_{j_\ell}w_{j_\ell,j_{\ell+1}}\sqrt{\sum_{j_1,\dots,j_{\ell-1}}w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}}.$$

The full story, however, is not complete—these calculations only establish coverings with respect to the $\mathbb{L}^1$ metric, which, when converted to $\mathbb{L}^2$ minimax upper bounds as per [Yang and Barron, 1999], lead to suboptimal rates of the form $\left(\frac{L^3\mathscr{V}^2\log(\max_{\ell\ge 2}d_\ell)}{n}\right)^{1/3}$ with exponent (1/3) instead of (1/2). In order to obtain squared-error risk bounds of the form (7), we can adapt these accuracy bounds to the $\mathbb{L}^2$ metric. Such extensions require more finesse. For example, when expanding the second power of (11) and using the triangle inequality, one is led to bound terms of the form

$$\sqrt{\mathrm{E}[|\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|^2|K_{j_\ell}]}, \tag{15}$$

which do not have the same form as (12). Indeed, the conditional second moments of $\widetilde{w}_{j_{\ell-1},j_\ell}$ are biased upwards, since each $K_{j_{\ell-1},j_\ell}$ is conditionally binomially distributed $\mathrm{Bin}(K_{j_\ell}, p_{j_{\ell-1}|j_\ell})$ given K_{j_ℓ}.[3] Thus,

$$\sqrt{\mathrm{E}[|\widetilde{w}_{j_1}\widetilde{w}_{j_1,j_2}\cdots\widetilde{w}_{j_{\ell-1},j_\ell}|^2|K_{j_\ell}]} \gg w_{j_1}w_{j_1,j_2}\cdots w_{j_{\ell-1},j_\ell}\frac{K_{j_\ell}}{p_{j_\ell}M}. \tag{16}$$

Instead, one can introduce a small correction term to the $\widetilde{w}_{j_{\ell-1},j_\ell}$, without sacrificing approximation error, by truncating the $K_{j_{\ell-1},j_\ell}$ below level 2 and redefining $\widetilde{p}_{j_\ell|j_{\ell-1}} = \frac{(K_{j_{\ell-1},j_\ell}-1)}{K_{j_{\ell-1}}}\mathbb{1}_{\left\{K_{j_{\ell-1},j_\ell}\ge 2\right\}}$. This empirical quantity can also be decomposed as

$$\underbrace{\frac{K_{j_{\ell-1},j_\ell}}{K_{j_{\ell-1}}}}_{\text{unbiased estimate of } p_{j_\ell|j_{\ell-1}}} \quad - \quad \underbrace{\frac{\mathbb{1}_{\left\{K_{j_{\ell-1},j_\ell}>0\right\}}}{K_{j_{\ell-1}}}}_{\text{correction}}. \tag{17}$$

Note that now $\widetilde{p}_{j_\ell|j_{\ell-1}}$ are empirical sub-probabilities, since they no longer sum to one across the indices j_ℓ. The first term of (17) is an unbiased estimate of $p_{j_\ell|j_{\ell-1}}$ and the second correction term is of smaller order. This truncation only further sparsifies the network by requiring that the size of the weights exceed some threshold value; otherwise they are set to zero. By

[3] For example, if $K \sim \mathrm{Bin}(K', p)$, then the second moment of K/K' is $p^2 + p(1-p)/K'$, whereas the second moment of $K\mathbb{1}_{\{K\ge 2\}}/K'$ is at most p^2.

(22) in Lemma 1, the redefined network weights yield the inequality $\mathrm{E}[|\phi(\widetilde{z}_{j_\ell}) - \phi(z_{j_\ell})||K_{j_\ell}] \leq 2v_{j_\ell}\frac{\sigma_{j_\ell}(\mathbf{x})}{\sqrt{K_{j_\ell}}}$.

Now, to the matter of unknown v_{j_ℓ}. Each one can be approximated by a distribution [independent of $\widetilde{\mathbf{W}}$] that puts positive mass on the two integer values that straddle it, namely,

$$\widetilde{v}_{j_\ell} = \frac{1}{M}[\lceil v_{j_\ell}\rceil (T_{j_\ell} - 1)\mathbb{1}_{\{T_{j_\ell} \geq 2\}} + \lfloor v_{j_\ell}\rfloor (M - T_{j_\ell} - 1)\mathbb{1}_{\{M - T_{j_\ell} \geq 2\}})],$$

where $T_{j_\ell} \sim \mathrm{Bin}(M, q_{j_\ell})$ are independent and $v_{j_\ell} = \lceil v_{j_\ell}\rceil q_{j_\ell} + \lfloor v_{j_\ell}\rfloor (1 - q_{j_\ell})$. By Lemma 1,

$$\mathrm{E}[|\widetilde{v}_{j_\ell} - v_{j_\ell}|^2] \leq \frac{2}{M}(\sqrt{q_{j_\ell}}\lceil v_{j_\ell}\rceil + \sqrt{1 - q_{j_\ell}}\lfloor v_{j_\ell}\rfloor)^2.$$

It can similarly be shown via Lemma 1 that the expected value of $\widetilde{v}^2_{j_\ell}$ is at most $(\lceil v_{j_\ell}\rceil q_{j_\ell} + \lfloor v_{j_\ell}\rfloor (1 - q_{j_\ell}))^2 = v^2_{j_\ell}$. To summarize, if $\widetilde{w}_{j_{\ell-1},j_\ell} = \widetilde{v}_{j_{\ell-1}}\widetilde{p}_{j_\ell|j_{\ell-1}}$, then the square root of the conditional expected second moment of their successive products (15) is at most (12).

Realizations of $\widetilde{v}_{j_\ell}$ can be indexed by triplets of integers [corresponding to $\lceil v_{j_\ell}\rceil$, $\lfloor v_{j_\ell}\rfloor$, and T_{j_ℓ}] and, as such, they can be shown to have manageable cardinality as well. In fact, only $LM/2$ of the v_{j_ℓ} need to be represented, since the total number of nonzero weights $\widetilde{p}_{j_\ell|j_{\ell-1}}$ in $f(\mathbf{x}; \widetilde{\mathbf{W}})$ is at most $LM/2$. Indeed, the counts $K_{j_{\ell-1},j_\ell}$ which determine the weights $\widetilde{p}_{j_\ell|j_{\ell-1}}$ have sum $\sum_{j_{\ell-1},j_\ell} K_{j_{\ell-1},j_\ell} = M$. Accordingly, the number of nonzero weights from layer $\ell - 1$ to layer ℓ, namely $\sum_{j_{\ell-1},j_\ell} \mathbb{1}_{\{K_{j_{\ell-1},j_\ell} \geq 2\}}$, is not more than $M/2$ and hence the total number of nonzero weights is not more than $LM/2$. These arise from M choices of nonzero weight paths $(j_1, j_2, \dots, j_L)$ from the $K_{j_1,j_2,\dots,j_L}$.

Define $\|\mathbf{W}\|_{1,\infty} = \max_\ell \|\mathbf{W}_\ell\|_{1,\infty}$, the maximum row-sums of the weight matrices. Then the triplets $(\lceil v_{j_\ell}\rceil, \lfloor v_{j_\ell}\rfloor, T_{j_\ell})$ can be indexed by a set with cardinality at most $[2(\|\mathbf{W}\|_{1,\infty} + 1)(M+1)]^{\lceil LM/2\rceil}$. Combining these counts with (10), it follows that if $\|\mathbf{W}\|_{1,\infty} \leq \mathscr{W}$, then the total cardinality of the representer set is at most

$$[2(\mathscr{W} + 1)(M+1)]^{\lceil LM/2\rceil}\binom{d_1 d_2 \cdots d_L + M - 1}{M}, \tag{18}$$

or, in other words, of order $LM\log(\mathscr{W}\max_{\ell\geq 1} d_\ell)$. This establishes the following statement about the accuracy and cardinality of a sparse covering of deep neural networks with general activation function [c.f., Theorem 1].

Theorem 2. *Let ϕ be a Lipschitz(1) activation function with $\phi(0) = 0$. Consider the parameterized family $\mathcal{F}(L, \mathscr{V}, \mathscr{W})$ of depth L networks with average path variation V at most $\mathscr{V}$ and $\|\mathbf{W}\|_{1,\infty} \leq \mathscr{W}$. There is a subfamily $\widetilde{\mathcal{F}}_M$ with log-cardinality of order*

$$LM\log(\mathscr{W}\max_{\ell\geq 1} d_\ell)$$

such that for any probability measure $\mathbb{P}$ on $[-1,1]^d$ and any $f(\mathbf{x}; \mathbf{W})$ belonging to $\mathcal{F}(L, \mathscr{V}, \mathscr{W})$, there is a sparse approximant $f(\mathbf{x}; \widetilde{\mathbf{W}})$ in $\widetilde{\mathcal{F}}_M$, with at most LM nonzero weights, such that

$$\int |f(\mathbf{x}; \widetilde{\mathbf{W}}) - f(\mathbf{x}; \mathbf{W})|^2 \mathbb{P}(d\mathbf{x}) \leq \left[\frac{L\mathscr{V}}{\sqrt{M}}\right]^2.$$

Remark 2. *One can also create a cover with cleaner cardinality via an alternative argument. To do this, assume that there are reference weights* $\mathbf{W}^* = (\mathbf{W}_1^*, \dots, \mathbf{W}_L^*)$ *such that* $v_{j_\ell} \le v_{j_\ell}^*$*, where* $v_{j_\ell}^* = \sum_{j_{\ell+1}} w_{j_\ell, j_{\ell+1}}^*$*. Then, at each layer, arrange a null index* j_ℓ^0 *for* j_ℓ *with corresponding trivial subnetwork* $f_{j_\ell^0}(x) = z_{j_\ell^0} = 0$ *and define* $p_{j_\ell | j_{\ell-1}}^* = \frac{w_{j_\ell, j_{\ell-1}}}{v_{j_{\ell-1}}^*}$ *and* $p_{j_\ell^0 | j_{\ell-1}}^* = 1 - \frac{v_{j_{\ell-1}}}{v_{j_{\ell-1}}^*}$*. Each unit in the representer network* $f(\mathbf{x}; \widetilde{\mathbf{W}})$ *computes* $x_{j_{\ell-1}} = \phi\big(\sum_{j_\ell} x_{j_\ell} \widetilde{w}_{j_{\ell-1}, j_\ell}\big)$*, with* $\widetilde{w}_{j_{\ell-1}, j_\ell} = v_{j_{\ell-1}}^* \widetilde{p}_{j_\ell | j_{\ell-1}}$ *and* $\widetilde{p}_{j_\ell | j_{\ell-1}}$ *are the empirical conditional probabilities built from counts of random draws from* $p_{j_1, \dots, j_L}^* = p_{j_1}^* p_{j_2|j_1}^* \cdots p_{j_L | j_{L-1}}^*$*. Adding the null index increases the interlayer dimension* d_ℓ *by one, so redefining* $d_\ell + 1$ *by* d_ℓ*, we find that the cardinality of the representer set is at most*

$$\binom{d_1 d_2 \cdots d_L + M - 1}{M}.$$

The difference now is that the approximation accuracy has changed slightly. In particular, the average path variation V *becomes*

$$V^* = \frac{1}{L} \sum_{\ell=0}^{L-1} \sum_{j_1, j_2, \dots, j_\ell, j_{\ell+1}} w_{j_1} w_{j_1, j_2} \cdots w_{j_{\ell-1} j_\ell} w'_{j_\ell, j_{\ell+1}} = \frac{1}{L} \sum_{\ell=0}^{L-1} \|\mathbf{W}_1 \mathbf{W}_2 \cdots \mathbf{W}_\ell \mathbf{W}'_{\ell+1}\|_1, \quad (19)$$

where $w'_{j_\ell, j_{\ell+1}} = w_{j_\ell, j_{\ell+1}}^* \sigma_{j_\ell}^* \sqrt{1/p_{j_\ell}^*}$ *and with similar definitions for* $\sigma_{j_\ell}^*$ *and* $p_{j_\ell}^*$ *as before.*

Remark 3. *The complexity constant (14) is similar to the complexity constant* $V = \|\mathbf{W}_1 \mathbf{W}_2 \cdots \mathbf{W}_L\|_1$ *for linear (5) or ReLU networks [see Theorem 1] in that it involves only norms of products of the weight matrices.*

4 Adaptive Estimation

Flexible regression models are built by combining simple functional forms, which here consist of repeated compositions and linear transformations of nonlinear functions. In fitting such models to data in a training sample, there is a role for empirical performance criteria such as penalized squared error in selecting components of the function from a given library of candidate terms. With suitable penalty, optimizing the criterion adapts the total weights $(\mathbf{W}_1, \dots, \mathbf{W}_L)$ of combination or the number d_ℓ of units x_{j_ℓ} as well as the subset of which units to include in each layer. In practice, one does not know the "true" $V(f^*)$ for the regression function f^*, which makes it difficult to select an upper bound v on $V = V(f)$ for functions f in $\mathcal{F}(L, \mathscr{V})$. In fact, f^* may not even be equal to a deep neural network and therefore empirical risk minimization over a finite covering of $\mathcal{F}(L, \mathscr{V})$ is inconceivable unless the model is well-specified.

Motivated by the previous concerns, an important question is whether the same rate in (7), derived from nonadaptive estimators, is available from an adaptive risk bound (3) [which allows for a more data-dependent and agnostic criterion for fits of f^*] for estimators that minimize a penalized empirical risk (2) over $\mathcal{F}(L) = \bigcup_{\mathscr{V} > 0} \mathcal{F}(L, \mathscr{V})$.

We investigate penalized estimators $\hat{f}$ with penalty defined through the "smallest" complexity constant $V(f^*)$ [see (14)] among all representations of a network f^*, i.e.,

$$\hat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}(L)} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2 + \lambda_n V(f) \right\},$$

where $\lambda_n \asymp \left(\frac{L^3 \log(\max_{\ell \geq 2} d_\ell)}{n}\right)^{1/2}$ [whose choice is inspired by Theorem 1 and (7)]. Using techniques from [Klusowski and Barron, 2016], it can be shown that $\hat{f}$ has an adaptive risk bound of the form

$$\mathbb{E}\left[\|f^* - \hat{f}\|^2\right] \leq C \inf_{f \in \mathcal{F}(L)} \left\{\|f - f^*\|^2 + \lambda_n V(f)\right\}, \tag{20}$$

for some universal constant $C > 1$, which expresses the approximation and complexity tradeoff in the same spirit as (3).

Penalties of similar flavor have already been established for single hidden layer networks, corresponding to $L = 2$. For example, using approximation results from [Barron, 1993], it was shown in [Barron, 1994] that (20) holds for a penalty $\lambda_n \|\mathbf{W}_1\|_1$ with $\lambda_n \asymp (d(\log(n/d))/n)^{1/2}$. Furthermore, recent work [Klusowski and Barron, 2016] has shown that one can additionally control the size of the hidden layer parameters w_{j_1,j_2} through a penalty $\lambda_n \|\mathbf{W}_1 \mathbf{W}_2\|_1$ with $\lambda_n \asymp ((\log(d))/n)^{1/2}$ (which yields risk bounds similar to those for high-dimensional deep ReLU networks mentioned previously (7)). The authors' past work [Barron and Klusowski, 2018, Theorem 4] shows that these risk bounds for the single hidden layer case are essentially optimal in the sense that the minimax risk $R_n(\mathcal{F}(L, \mathscr{V}))$ is lower bounded by $\mathscr{V}^{1/4}((\log(d))/n)^{1/2}$.

The idea for establishing the validity of (20) builds on earlier work [Barron, 1991], which takes the penalty, penalty(f), to be Kraft summable, $\sum_{f \in \mathcal{N}_{\mathcal{F}(L,\mathscr{V})}(\epsilon_n)} e^{-\text{penalty}(f)} \leq 1$, where $\mathcal{N}_{\mathcal{F}(L,\mathscr{V})}(\epsilon_n)$ is a finite ϵ_n-covering of $\mathcal{F}(L, \mathscr{V})$. In this way, penalty$(f)$ is interpretable as a complexity [in nats] or $e^{-\text{penalty}(f)}$ is interpretable as a prior probability of f. For instance, one may assign penalty(f) to be the minimal log-cardinality of coverings of function classes, plus a description length of such classes, which in this case, is the codelength [in nats] to describe the subset of nonzero weights in f. These considerations motivate the choice of penalty as being proportional to the log-size of a sparse covering of deep networks in $\mathcal{F}(L, \mathscr{V})$ that achieve a desired accuracy [see (10)]. Using an accuracy quantification similar to Theorem 1 or Proposition 2, the choice of M can then be optimized to balance approximation error and complexity.

5 Future Work

One open question for future research involves network quantization and memory requirements for storing the network topology and the associated network weights. The previous approximation scheme of drawing M random indices $(j_1, j_2, \ldots, j_L)$ from the distribution $p_{j_1,j_2,\ldots,j_L}$ may be well suited for this purpose. For example, if in addition to establishing bounds on the expected value of $f(\cdot; \widetilde{\mathbf{W}})$, we were able show high-probability statements of the form

$$\mathrm{P}\left[\int |f(\mathbf{x}; \mathbf{W}) - f(\mathbf{x}; \widetilde{\mathbf{W}})|^2 \mathbb{P}(d\mathbf{x}) \leq C \log(1/\delta) L^2 V^2(f)/M\right] \geq 1 - \delta, \quad \delta \in (0, 1),$$

one could construct a quantization of the network from the aforementioned sampling scheme with high-probability. The bottleneck in such a quantization is from generating the empirical counts $K_{j_1,j_2,\ldots,j_L} \sim \text{Multinomial}(M, (p_{j_1,j_2,\ldots,j_L}))$, which are then marginalized to form the empirical network weights. It is possible to implement this random construction in $O(M(d_1 + d_2 + \cdots + d_L))$ time by sampling M random paths via L successive random draws from the conditionals $p_{j_\ell | j_{\ell-1}}$ [each of which is computable in $O(d_\ell)$ time].

Lemma 1. *Suppose $K \sim \mathrm{Bin}(m, p)$. Then,*

$$\mathbb{E}\left[(K-1)^2 \mathbb{1}_{\{K \geq 2\}}\right] \leq m(m-1)p^2, \tag{21}$$

$$\mathbb{E}\left[|(K-1)\mathbb{1}_{\{K \geq 2\}} - mp|^2\right] \leq 2mp, \tag{22}$$

and for any positive v_1 and v_2,

$$\begin{aligned} &\mathbb{E}\left[|v_1(K-1)\mathbb{1}_{\{K \geq 2\}} + v_2(m-K-1)\mathbb{1}_{\{m-K \geq 2\}}|^2\right] \\ &\quad \leq m(m-1)(v_1 p + v_2(1-p))^2 \end{aligned} \tag{23}$$

Proof. For showing (21), consider first the pointwise inequality $(k-1)^2 \mathbb{1}_{\{k \geq 2\}} \leq k(k-1)$, which is valid for any integer $k = 0, 1, \ldots, m$. Then take expectations with respect to K and note that $\mathbb{E}[K(K-1)] = m(m-1)p^2$.

For the second inequality (22), we note that $(K-1)\mathbb{1}_{\{K \geq 2\}} = K - \mathbb{1}_{\{K>0\}}$. Thus, it follows that

$$\mathbb{E}\left[|(K-1)\mathbb{1}_{\{K \geq 2\}} - mp|^2\right] \leq \mathsf{Var}[K] + \mathbb{E}\left[\mathbb{1}_{\{K>0\}}\right].$$

Finally, $\mathsf{Var}[K] = mp(1-p)$ and $\mathbb{E}\left[\mathbb{1}_{\{K>0\}}\right] = \mathbb{P}[K > 0] \leq mp$.

The third inequality (23) uses (21) together with the triangle inequality. □

References

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Andrew R Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics*, pages 561–576. Springer, 1991.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993. ISSN 0018-9448. doi: 10.1109/18.256500. URL `http://dx.doi.org/10.1109/18.256500`.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

Andrew R Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991. ISSN 0018-9448. doi: 10.1109/18.86996. URL `http://dx.doi.org/10.1109/18.86996`.

Andrew R. Barron and Jason M. Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

William Feller. An introduction to probability theory and its applications. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1971, 3rd ed.*, 1971.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In *Conference on Learning Theory*, pages 1064–1068, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Jason M. Klusowski and Andrew R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer, 2013.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

Gilles Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz")*, pages 1–12, 1980-1981. URL http://www.numdam.org/item/SAF_1980-1981____A5_0. talk:5.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2165799. URL http://dx.doi.org/10.1109/TIT.2011.2165799.

Philippe Rigollet and Alexandre B. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1–9. IEEE, 2015.

Yuhong Yang and Andrew R Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999. ISSN 0090-5364. doi: 10.1214/aos/1017939142. URL `http://dx.doi.org/10.1214/aos/1017939142`.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. *arXiv preprint arXiv:1802.03620*, 2018.